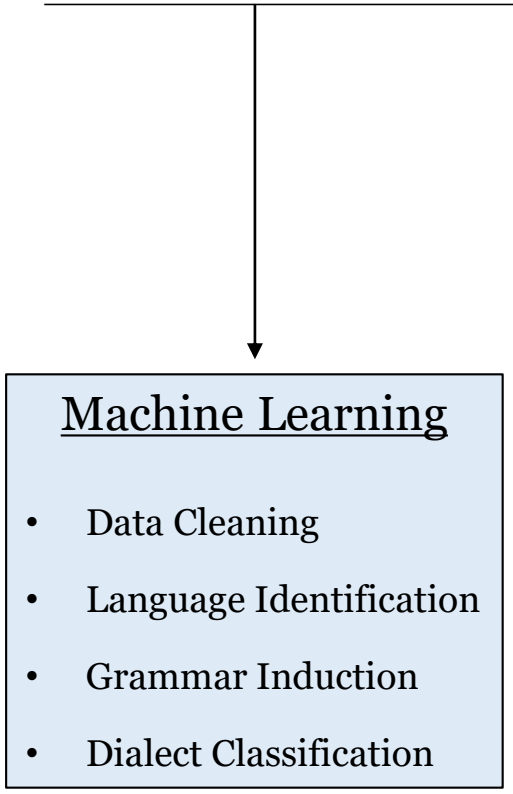
Traditional approaches require *asking* people once every decade what languages they use.

What if we could **speed up** and **scale up** this survey process?


Machine Learning
• Data Cleaning
• Language Identification
• Grammar Induction
• Dialect Classification

Machine Learning

- Data Cleaning
- Language Identification
- Grammar Induction
- Dialect Classification

DATA

	Twitter Corpus	Web Corpus
Africa, North	203,867,000	1,193,000,000
Africa, Southern	159,807,000	26,868,000
Africa, Sub	571,644,000	5,847,670,000
America, Brazil	156,705,000	2,265,007,000
America, Central	852,793,000	8,764,686,000
America, North	452,263,000	51,858,773,000
America, South	824,502,000	21,991,710,000
Asia, Central	220,106,000	26,940,518,000
Asia, East	198,177,000	19,229,783,000
Asia, South	580,221,000	15,087,017,000
Asia, Southeast	443,258,000	21,100,289,000
Europe, East	748,654,000	65,366,344,000
Europe, Russia	135,778,000	15,363,648,000
Europe, West	1,703,436,000	143,566,058,000
Middle East	421,926,000	1,721,259,000
Oceania	372,623,000	1,580,569,000
Total	**8.04 billion**	**401.90 billion**

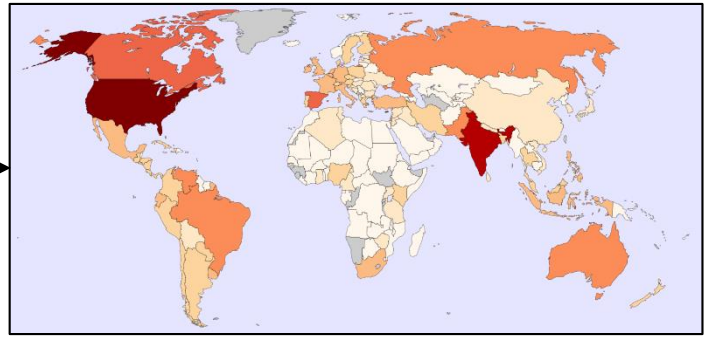
Machine Learning

- Data Cleaning
- Language Identification
- Grammar Induction
- Dialect Classification

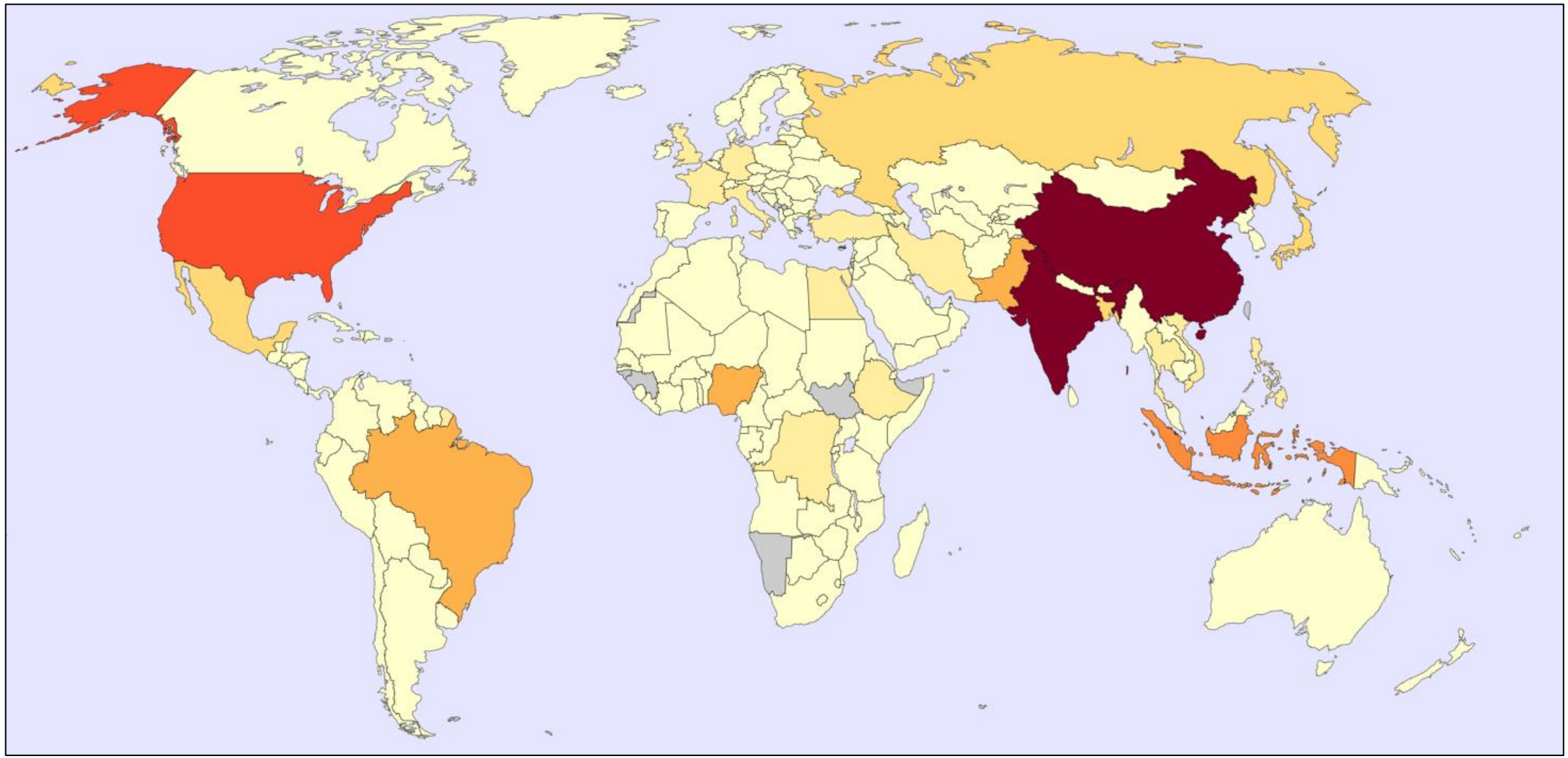
D A T A

	Twitter Corpus	Web Corpus
Africa, North	203,867,000	1,193,000,000
Africa, Southern	159,807,000	26,868,000
Africa, Sub	571,644,000	5,847,670,000
America, Brazil	156,705,000	2,265,007,000
America, Central	852,793,000	8,764,686,000
America, North	452,263,000	51,858,773,000
America, South	824,502,000	21,991,710,000
Asia, Central	220,106,000	26,940,518,000
Asia, East	198,177,000	19,229,783,000
Asia, South	580,221,000	15,087,017,000
Asia, Southeast	443,258,000	21,100,289,000
Europe, East	748,654,000	65,366,344,000
Europe, Russia	135,778,000	15,363,648,000
Europe, West	1,703,436,000	143,566,058,000
Middle East	421,926,000	1,721,259,000
Oceania	372,623,000	1,580,569,000
Total	**8.04 billion**	**401.90 billion**

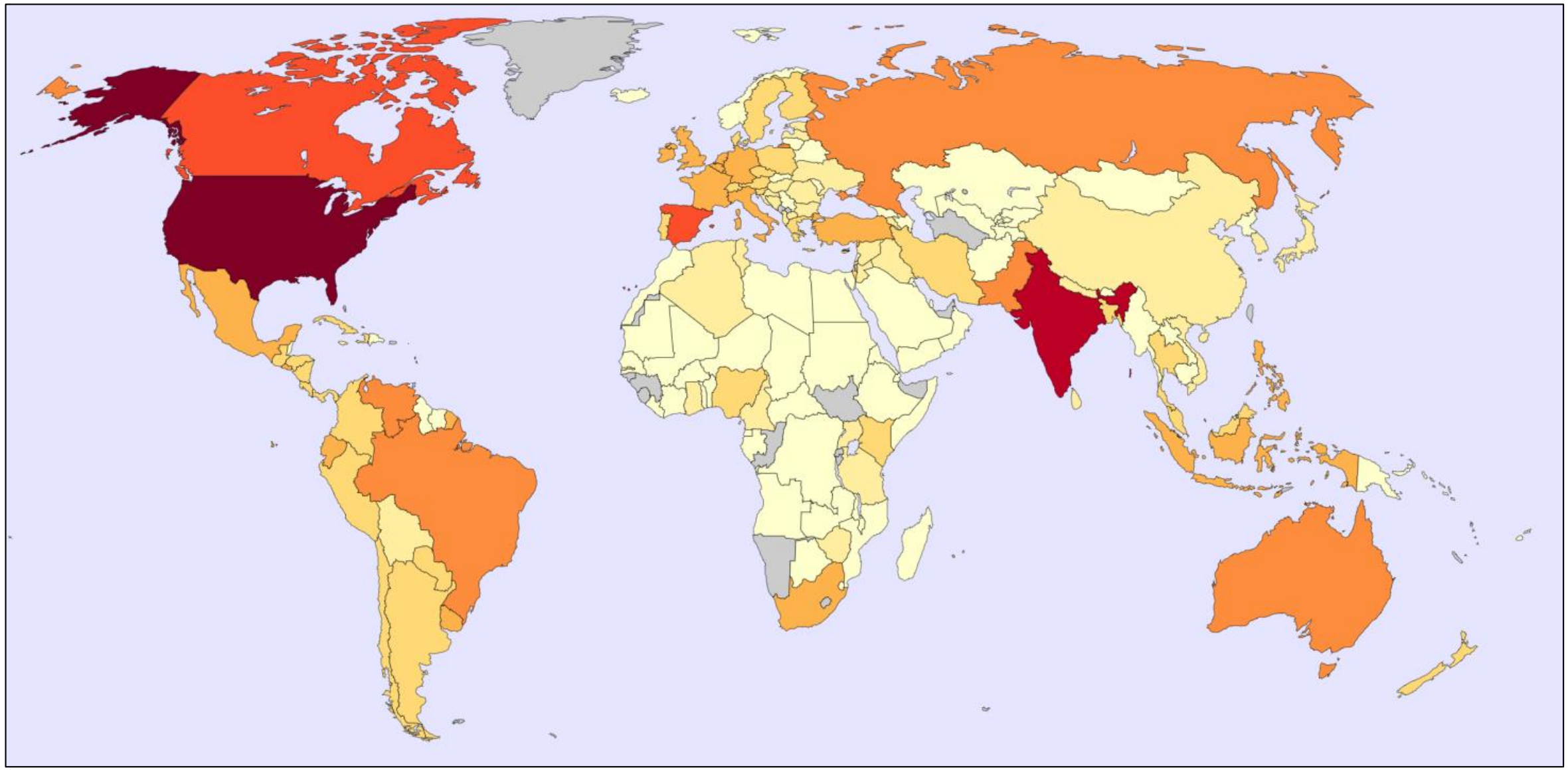
Language and Dialect Maps



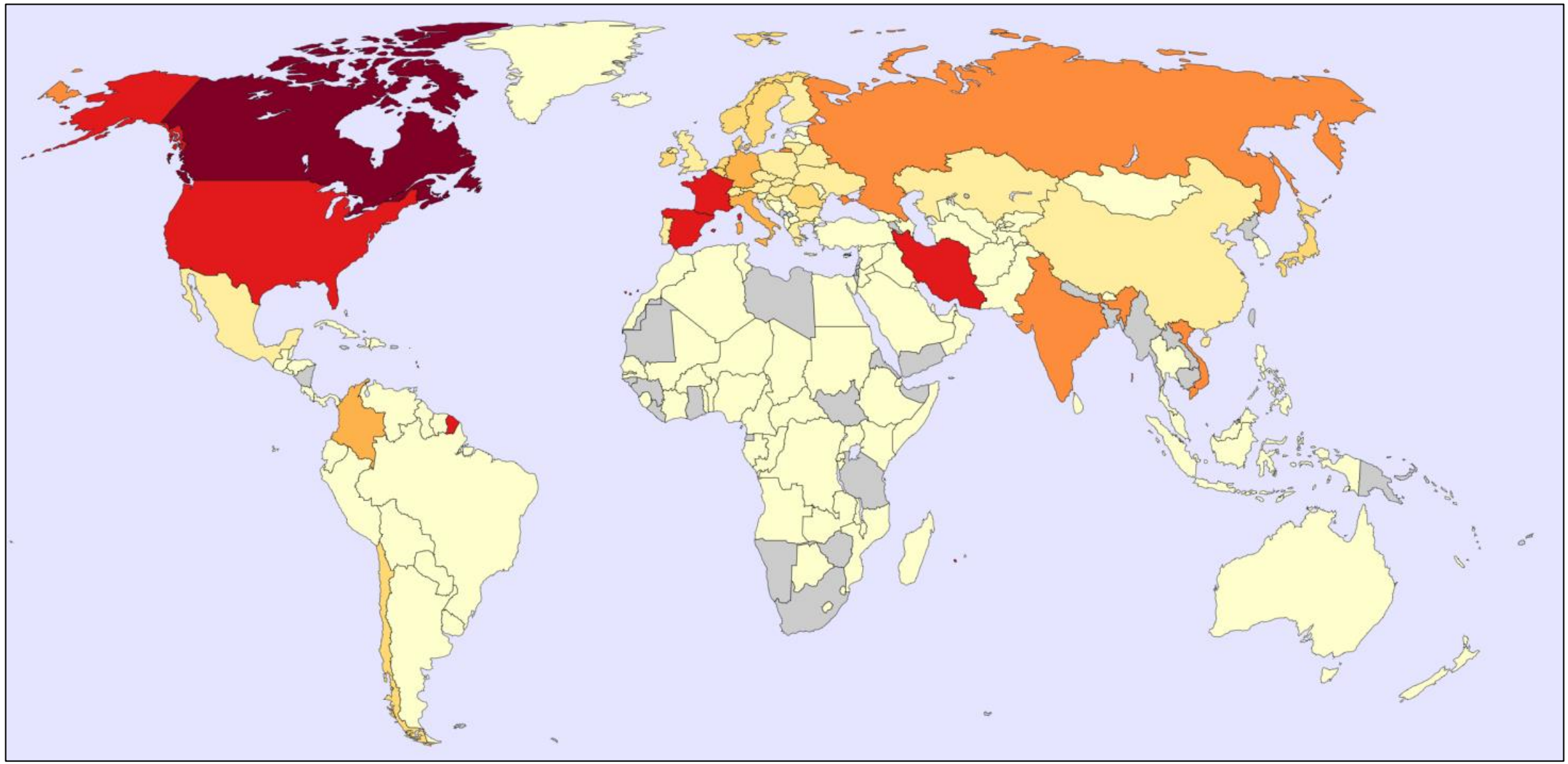
Population: Where do people live?



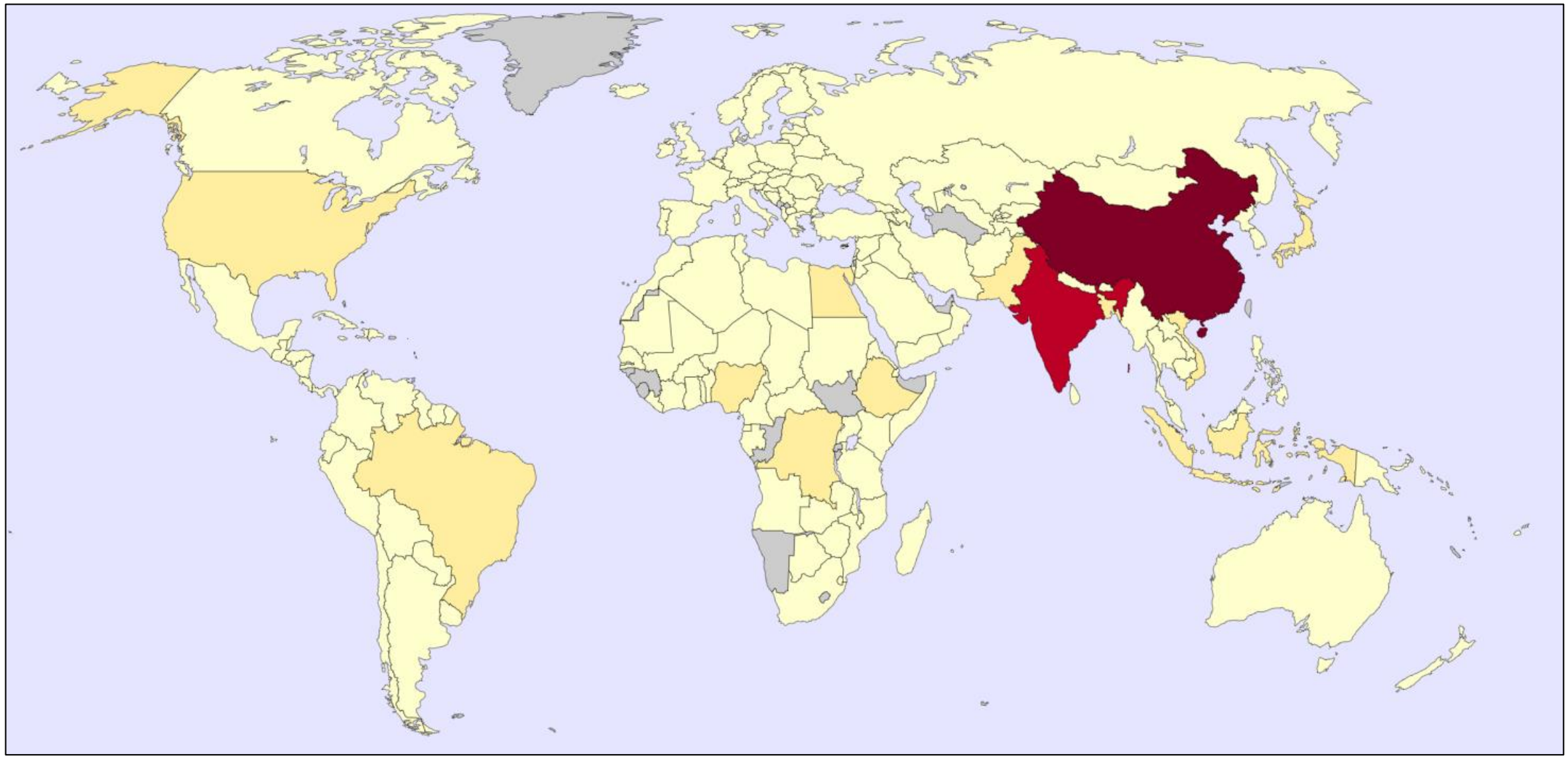
Twitter Density: Where do people Tweet?



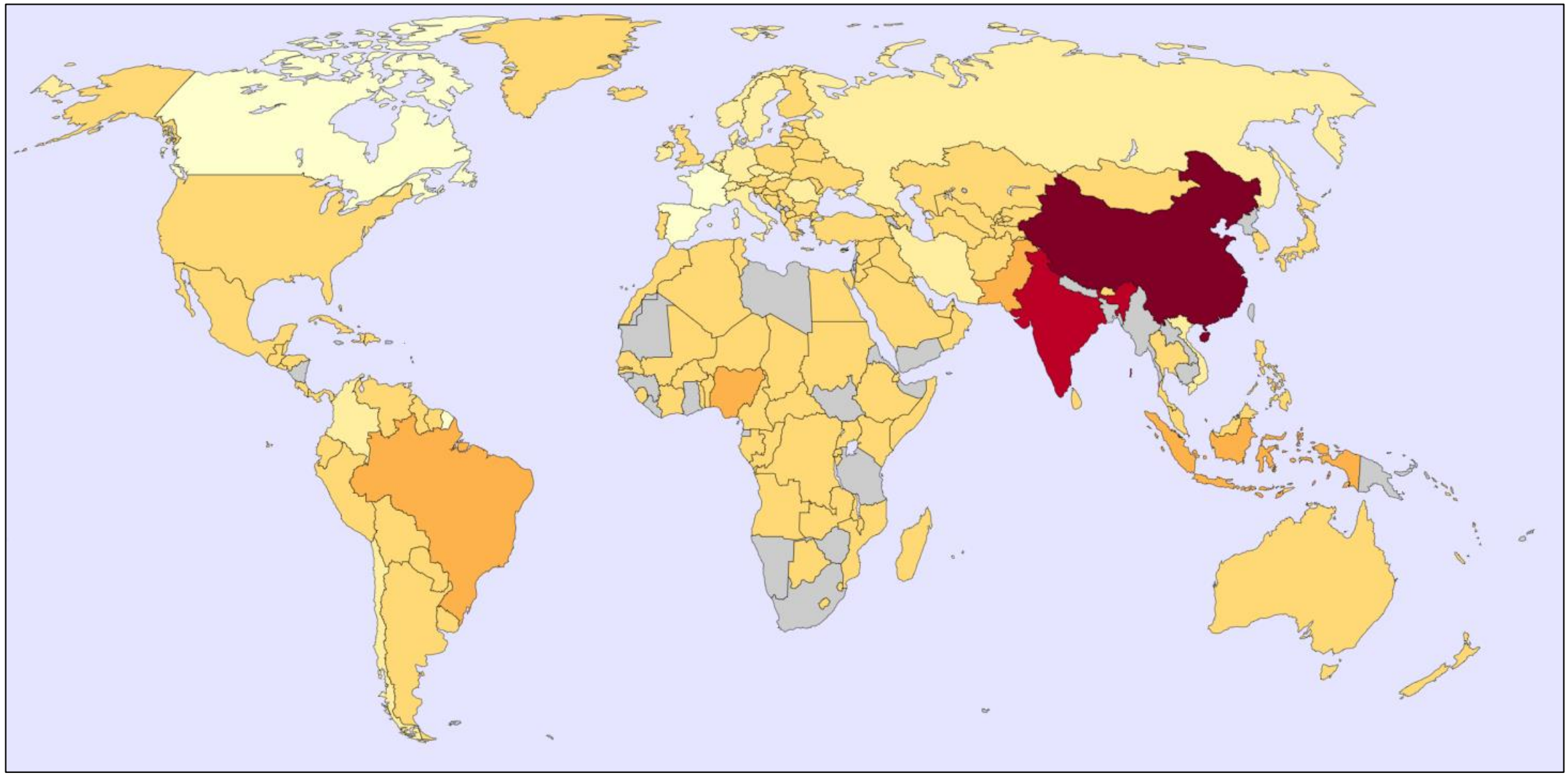
Web Density: Where do web documents come from?



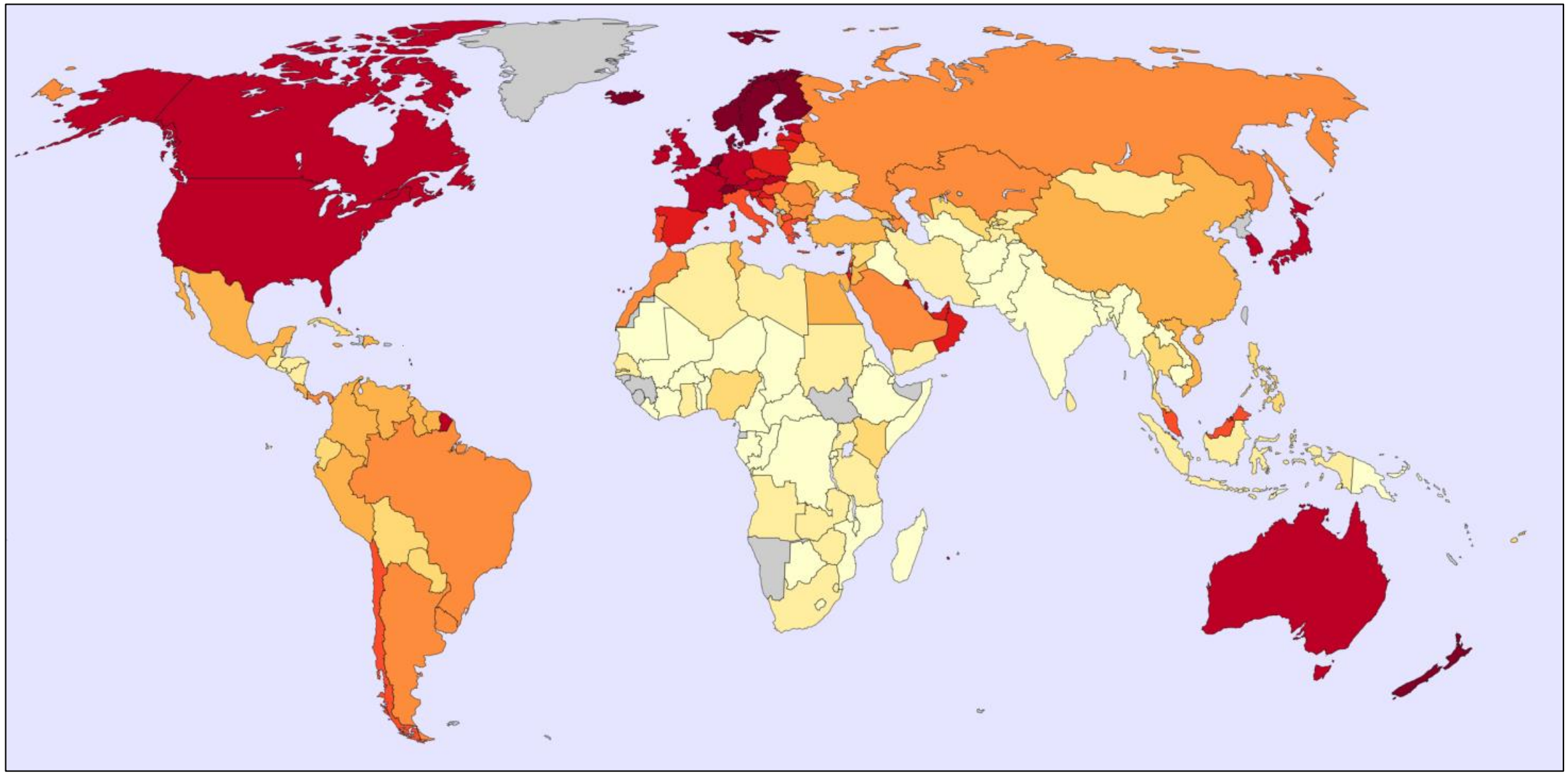
Where is Twitter not used? (i.e., under-represented countries)



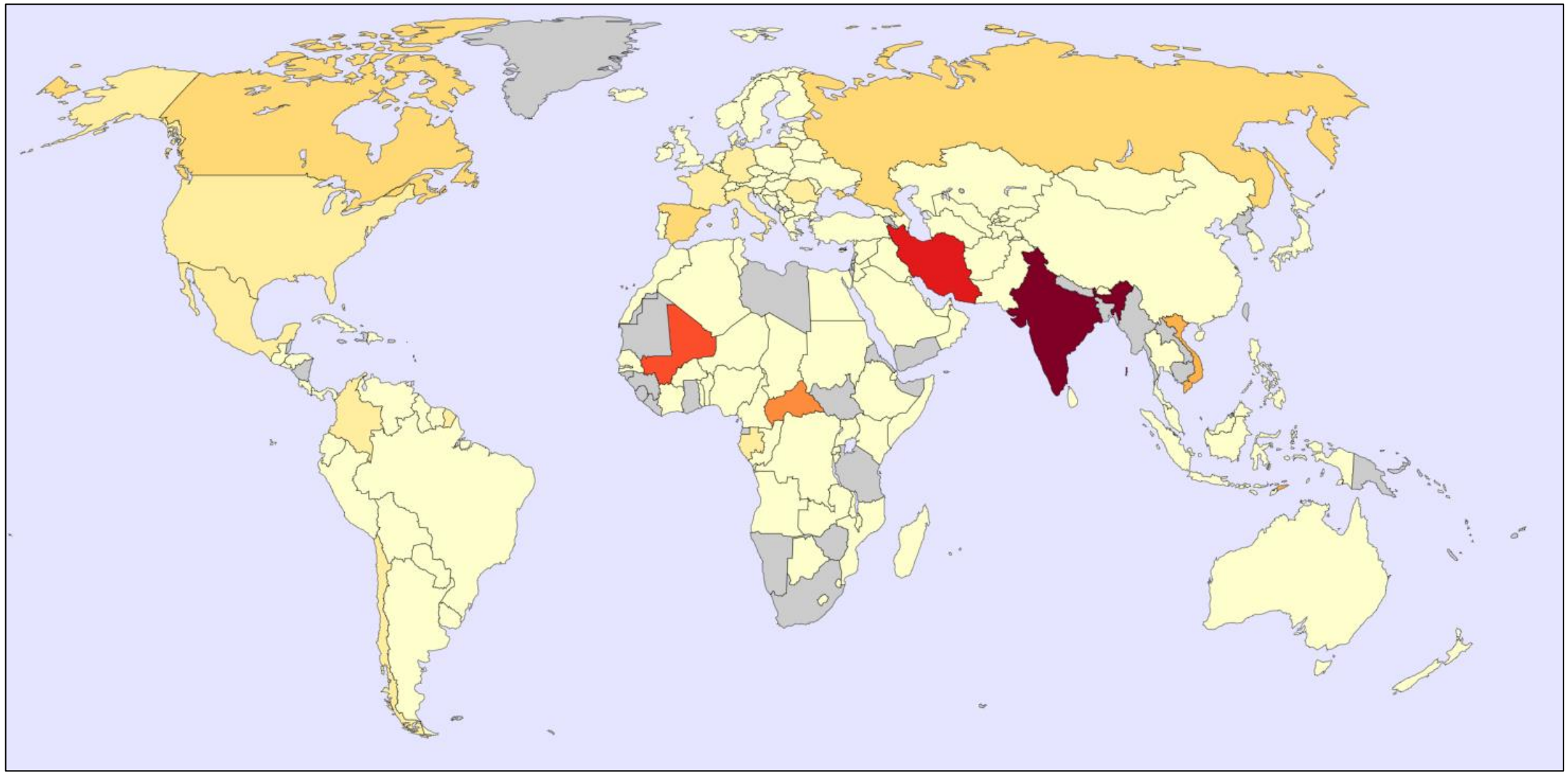
Where are web documents scarce? (i.e., under-represented countries)



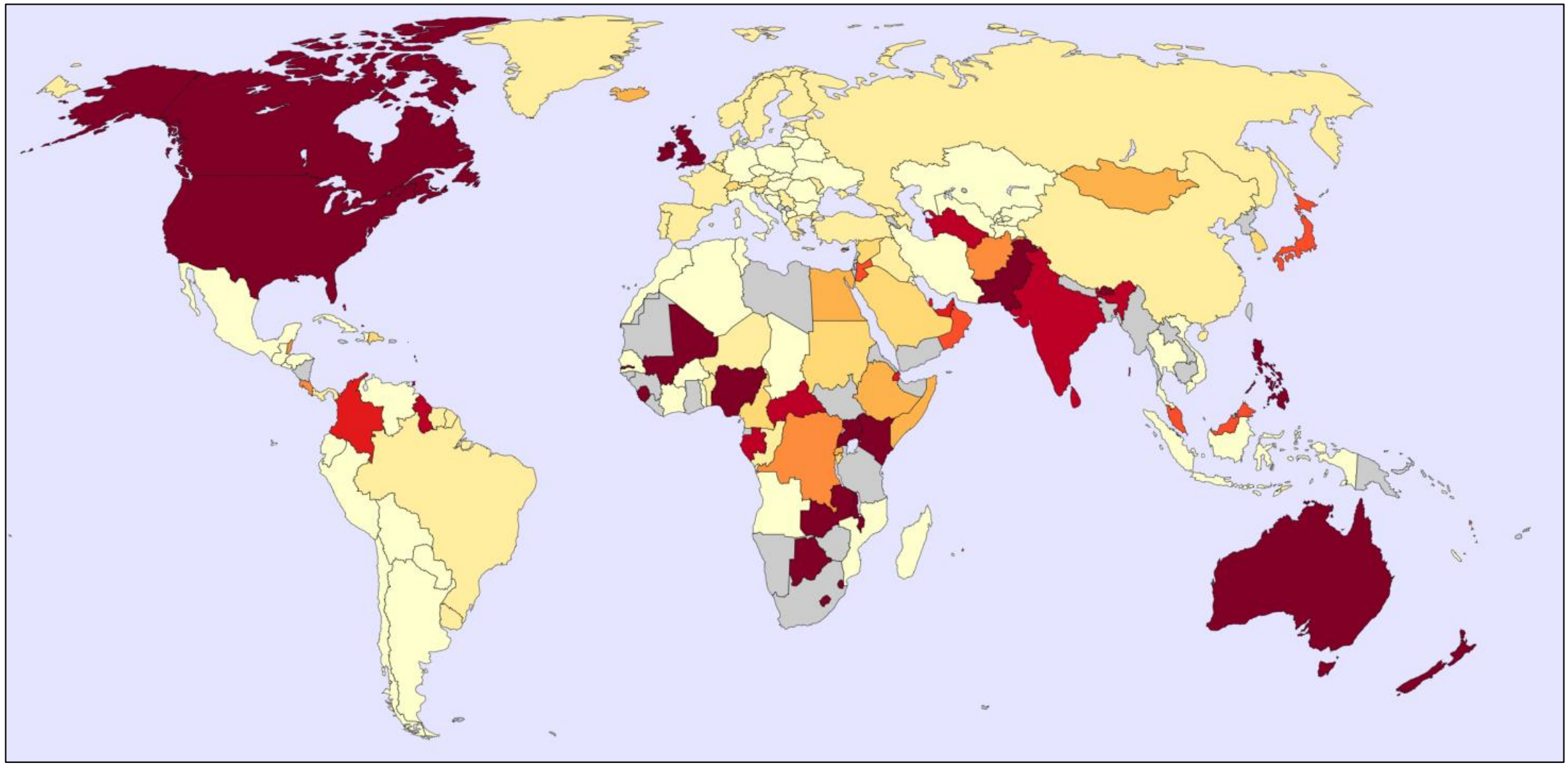
Internet Usage: Does the percentage of the population with internet access influence these datasets?



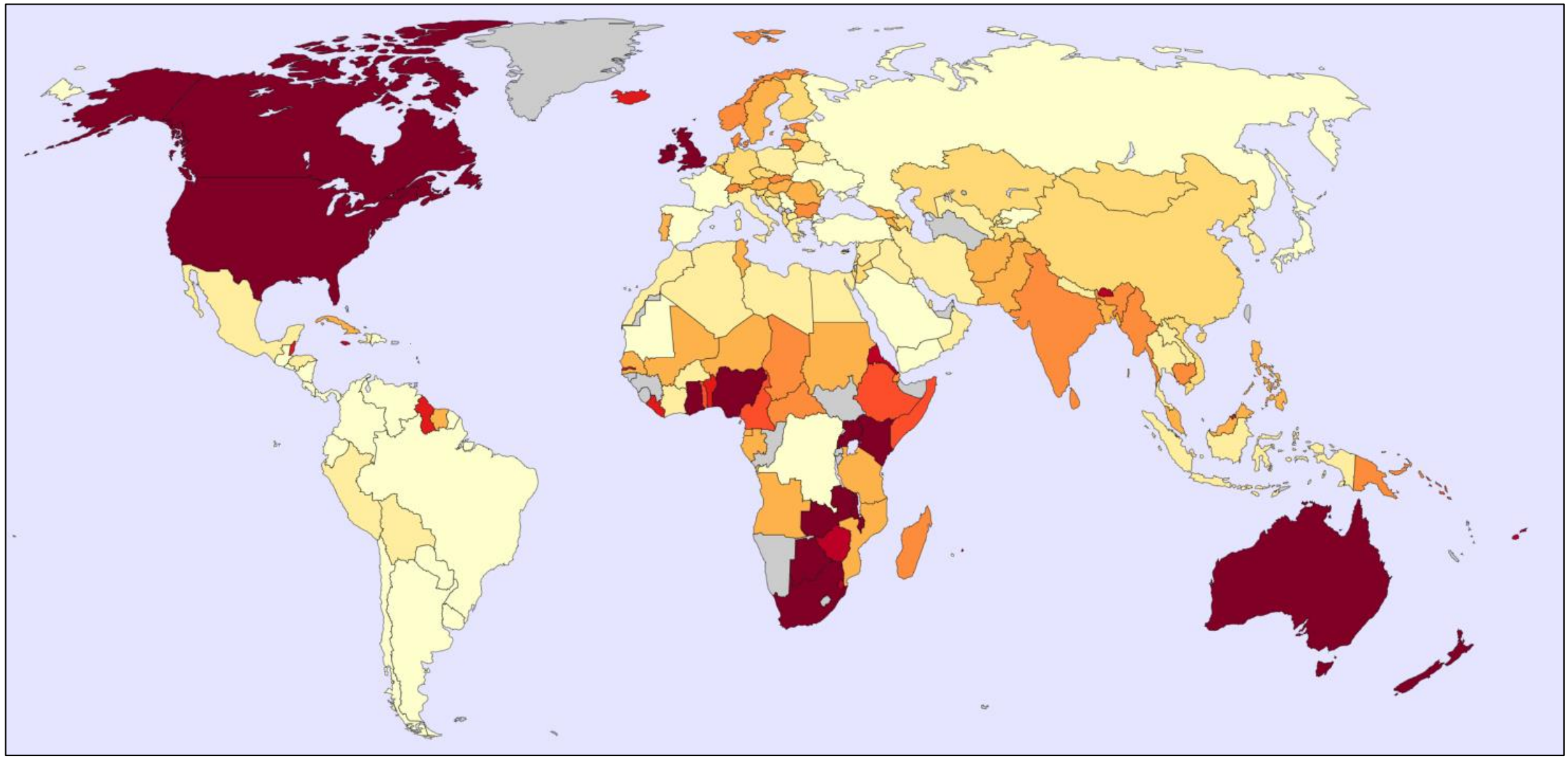
What would the web corpus look like if everyone had internet access?



Countries by their percent usage of English (web): Darker red means more monolingual usage

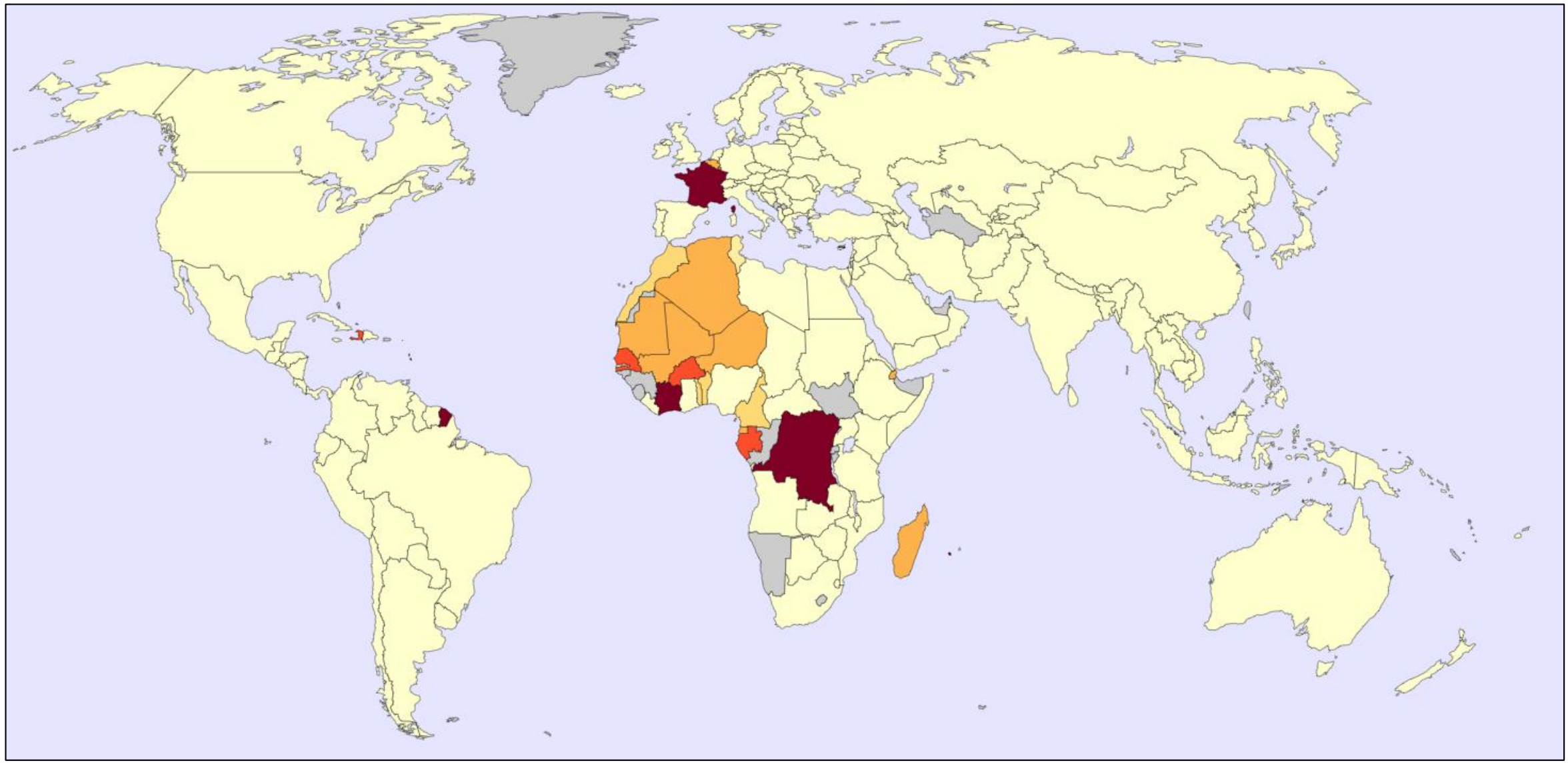


Countries by their percent usage of English (Twitter): Do the datasets capture different populations?

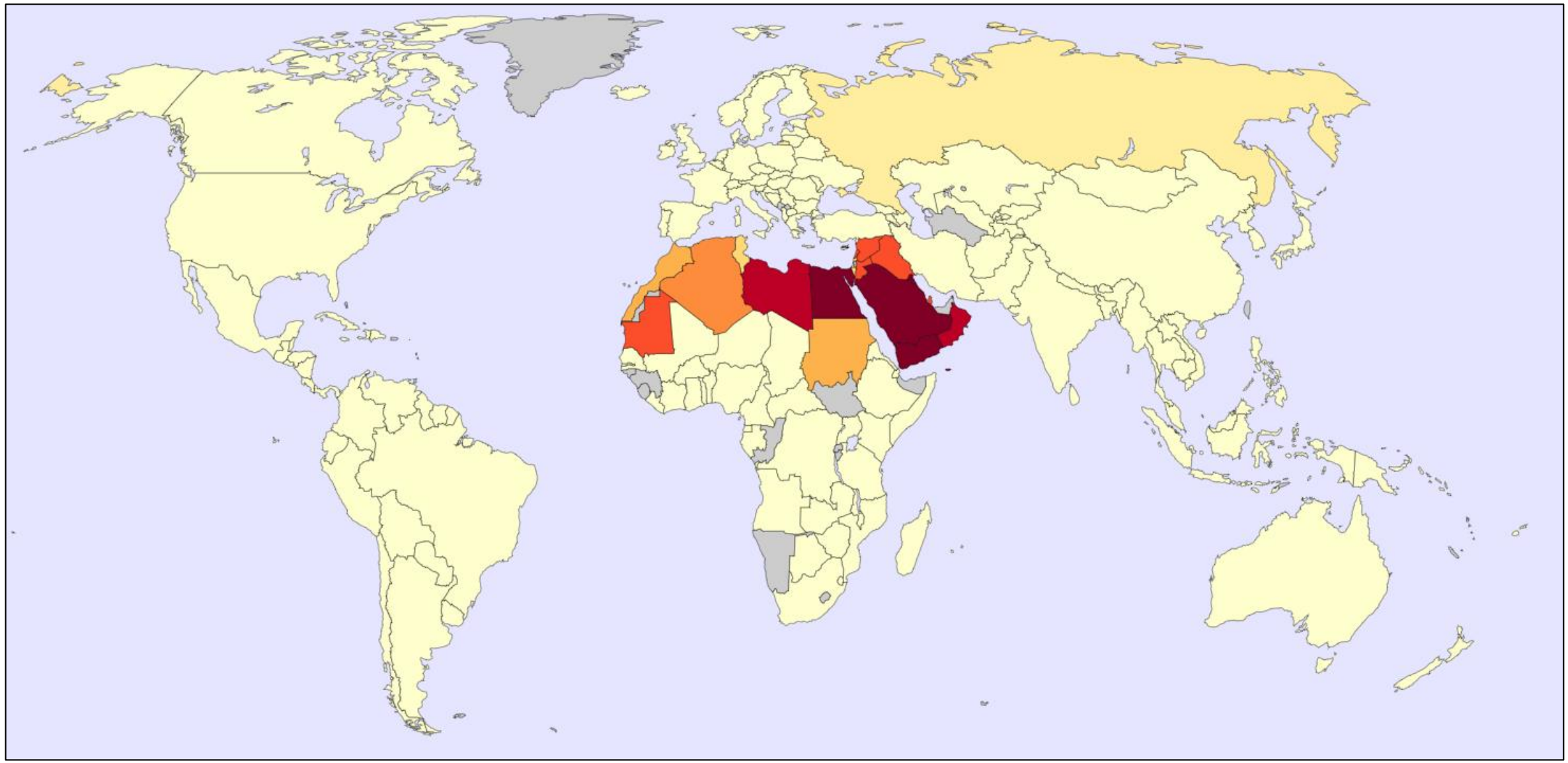


Countries by their percent usage of Spanish (Twitter): English is a global language, but not Spanish

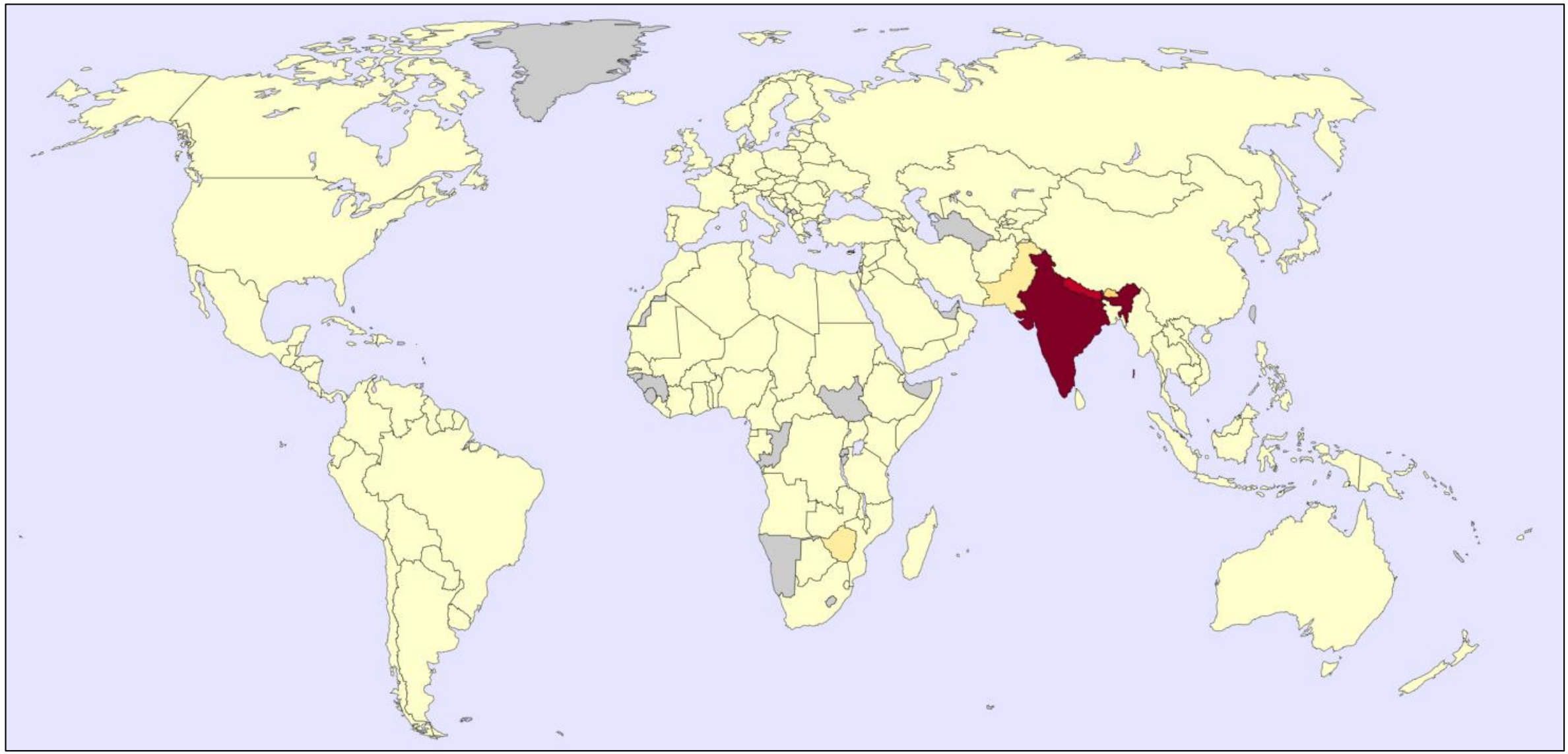
Countries by their percent usage of French (Twitter): French is also widely used



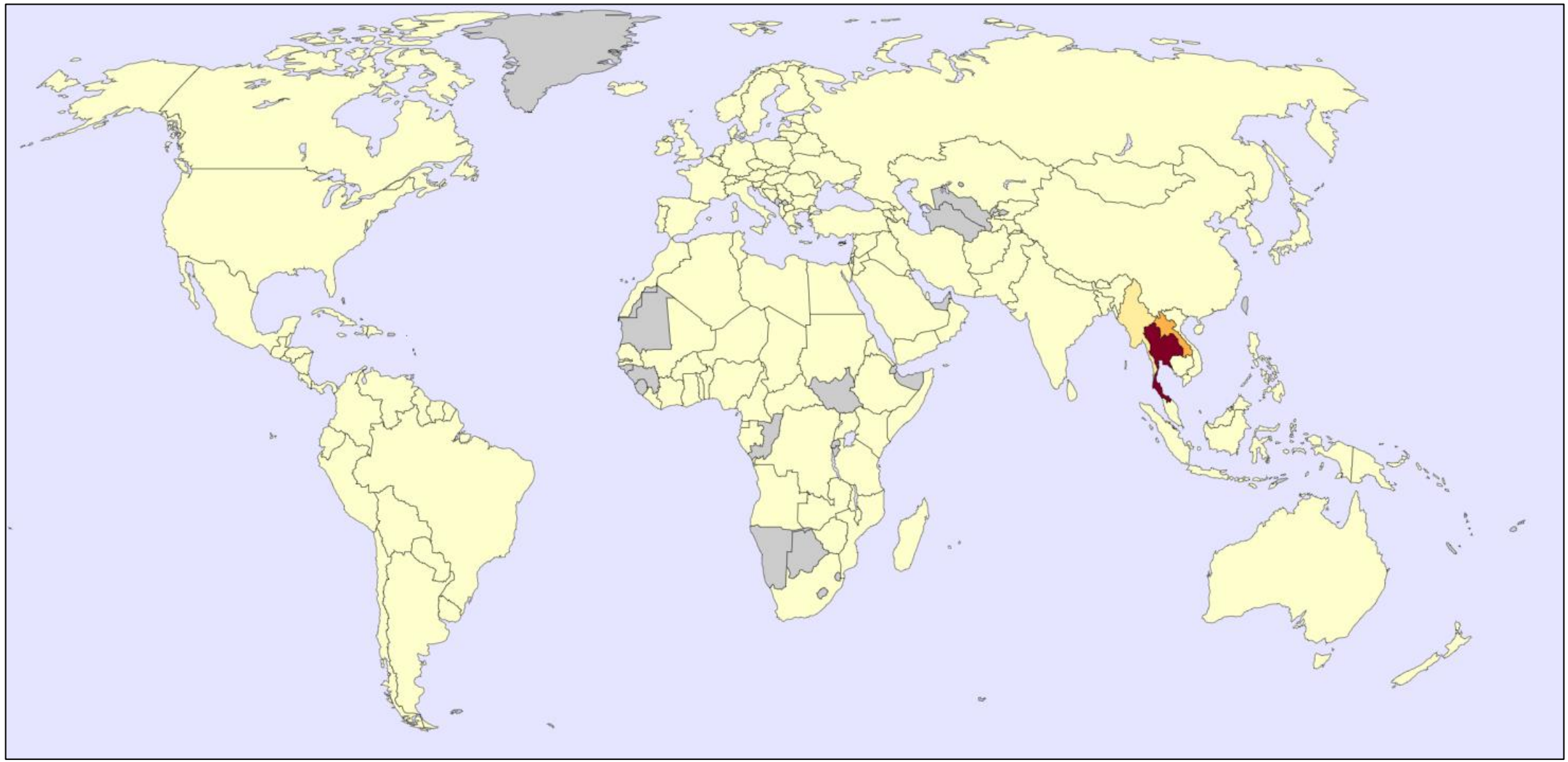
Countries by their percent usage of Arabic (Twitter): Some languages move with along moving populations



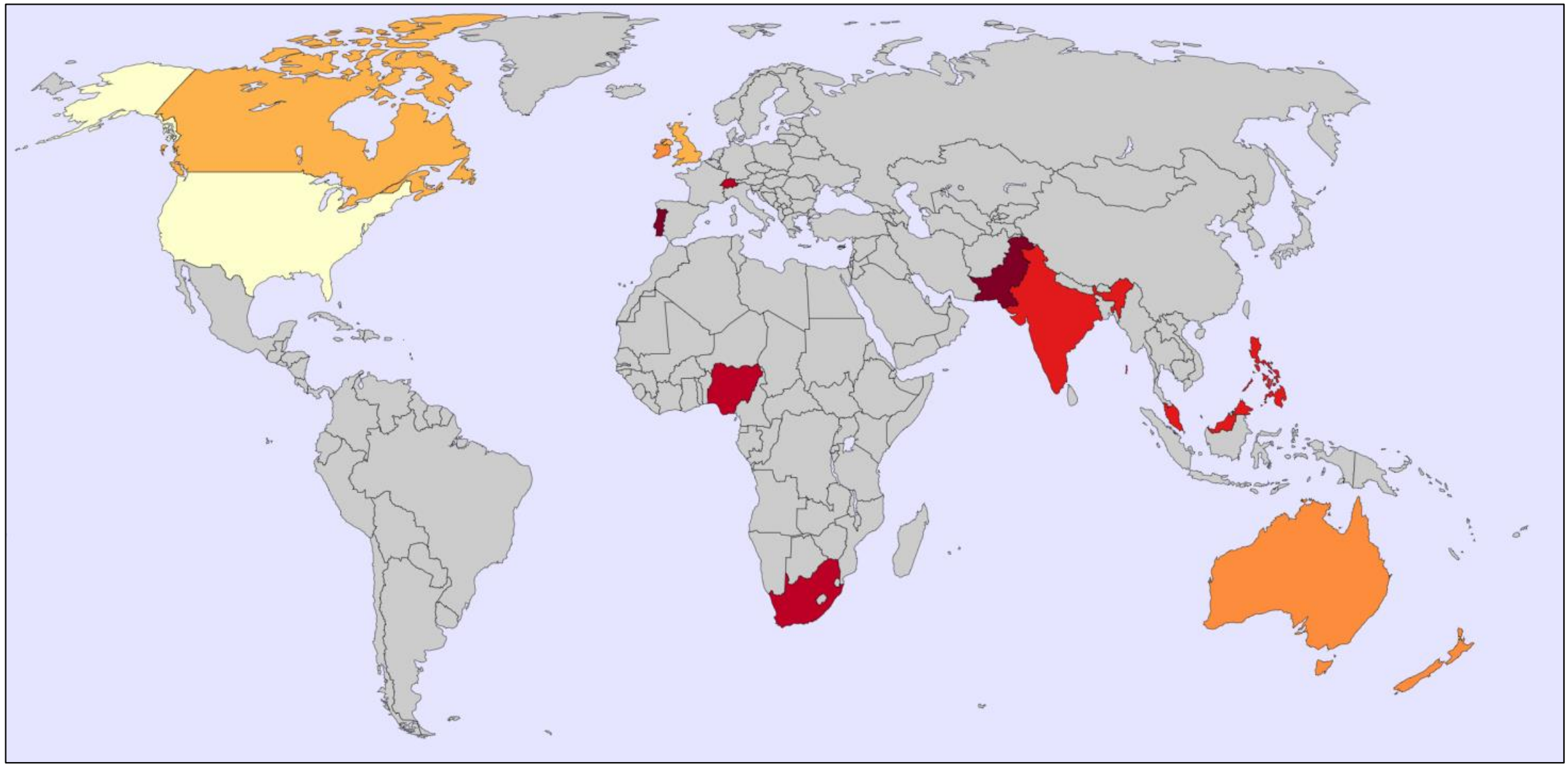
Countries by their percent usage of Hindi (Twitter): But others are restricted geographically



Countries by their percent usage of Thai (Twitter): Here, Thai is well-represented... but only in a narrow region

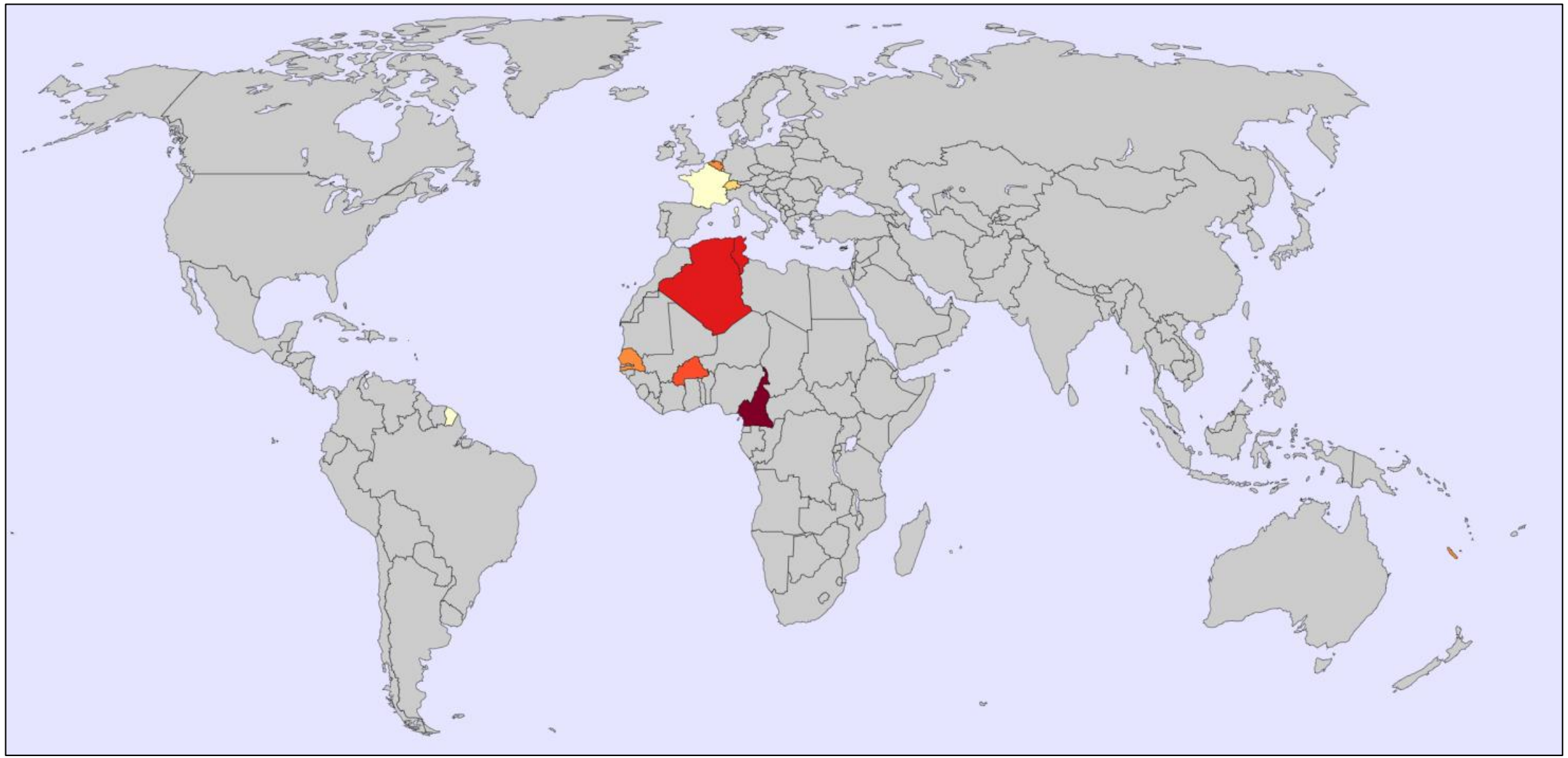


Uniqueness of English Dialects (Web): We can go beyond surveys by modelling the data from each country



Uniqueness of Spanish Dialects (Web): The dialect model shows the Spain is the central variety of Spanish

Uniqueness of French Dialects (Web): There aren't that many country-level dialects of French



Uniqueness of Russian Dialects (Web): Unlike other major languages, Russian is restricted to a (large) contiguous region

Explore the data at earthLings.io